Classification of five Chinese tea categories with different fermentation degrees using visible and near-infrared hyperspectral imaging

Jingming Ning, Jingjing Sun, Shuhuai Li, Mengge Sheng, and Zhengzhu Zhang
State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei, Anhui Province, China

ABSTRACT
A total of 206 samples of green, yellow, white, black, and Oolong teas were utilized to acquire hyperspectral imaging, and five tea categories were identified based on visible and near-infrared (NIR) hyperspectral imaging, combined with classification pattern recognition. The characteristic spectra were extracted from the region of interest (ROI), and the standard normal variate (SNV) method was preprocessed to reduce background noise. Four dominant wavelengths (589, 635, 670, and 783 nm) were selected by principal component analysis (PCA) as spectral features. Textural features were extracted by the Grey-level co-occurrence matrix (GLCM) from images at selected dominant wavelengths. Linear discriminant analysis (LDA), library support vector machine (Lib-SVM), and extreme learning machine (ELM) classification models were established based on full spectra, spectral features, textural features, and data fusion, respectively. Lib-SVM was the best model with the input data fusion, and the correct classification rate (CCR) achieved 98.39%. The results implied that visible and NIR hyperspectral imaging combined with Lib-SVM has the capability of rapidly and non-destructively classifying tea categories.

Introduction
Tea is one of the three most popular non-alcoholic beverages in the world, which contains a number of secondary metabolites, such as polyphenols, caffeine, theanine, and other components beneficial to human health.[1] In China, tea can be classified into six categories: green tea (unfermented), white tea (slightly fermented), yellow tea (partly fermented), Oolong tea (semi-fermented), black tea (fully fermented), and dark tea (post-fermented), according to the degree of fermentation.[2–4]

Tea consumption is expanding all over the world, but a layman will find it difficult to classify the different categories of teas. Different categories of tea have different qualities and quality standards. The classification of tea categories is important for consumers to purchase tea. Several chemical methods[5–8] were utilized to classify tea categories; however, these methods were time consuming and would destruct tea samples. Recently, green analytical techniques[9] revealed their domination in the classification of tea categories, which were fast, non-destructive, and precise, such as near-infrared (NIR) spectroscopy, computer vision, and electronic nose (E-nose). Chen et al.[10] utilized NIR spectroscopy combined with support vector machine (SVM) as recognition for the fast identification of green, black, and Oolong tea, and the identification accuracies were all above 90%. Wang et al.[11] proposed a computer-vision and machine-learning-based system to identify three tea categories. They used principal component analysis (PCA) to reduce the colour and texture...
information, which were fed into a fuzzy support vector machine (FSVM) with Winner-take-all (WTA) classifier, and they obtained an overall rate of 97.77%. Chen et al.\textsuperscript{[12]} developed a portable E-nose system based on the door imaging sensor array, and successfully used it for the classification of green tea, black tea, and Oolong tea. However, NIR spectroscopy just obtains single-point information of tea, computer vision only provides tea’s surface information, and E-nose merely captures the volatile organic compounds of tea. Hyperspectral imaging contains both the spectral and spatial information from the analysed sample, which has been recently used as emerging process analytical tool for food-quality control\textsuperscript{[13, 14]} Several researchers attempted to combine the spectral and spatial information, such as for differentiating fresh and frozen-thawed fish fillets,\textsuperscript{[15]} discriminating the varieties of black beans,\textsuperscript{[16]} quantifying the total volatile basic nitrogen content in chicken,\textsuperscript{[17]} and predicting the pH of salted pork\textsuperscript{[18]} wherein the models based on data fusion were better than those based on single variable matrix alone. However, there has been no research on the comprehensive classification of Chinese tea categories utilizing data fusion.

This study focused on the classification of five Chinese tea categories with different fermentation levels based on hyperspectral imaging. Spectral data were extracted within the range of 400–1000 nm wavelength from hyperspectral images of tea samples. Spectral and textural features were extracted and combined by level fusion. Linear discriminant analysis (LDA), library support vector machine (Lib-SVM), and extreme learning machine (ELM) models were developed to discriminate these five Chinese categories of teas.

**Materials and methods**

**Samples preparation**

Dark tea was always manufactured into brick, cake, and mushroom forms, which could easily be identified from the other teas. A total of 206 samples were collected from 13 provinces of China. Each tea category contained 44 samples, except yellow tea contained 30 samples. Each tea category had four samples to test the practicability of the final model. The remaining 186 samples were divided into two subsets in the ratio of 2:1, 124 samples were chosen as the calibration set, and the remaining 62 samples made up the prediction set. The calibration set was used to develop the model, while the prediction set was applied to test the robustness of the model.

**Hyperspectral image acquisition and processing**

The hyperspectral imaging system consisted of a hyperspectral spectrograph (Imspec torque V10E, Spectral Imaging Ltd., Oulu, Finland), two 150 W tungsten halogen lamps (3900, Illumination Technologies Inc., New York, USA), a moving platform, a dark box, and a computer with image acquisition and analysis software (Spectral Image Software, Isuzu Optics Corp., Taiwan, China). A total of 13 ± 0.5 g of sample was taken, and spread uniformly in the Petri dish (Φ×h: 9 cm × 1 cm). The Petri dish was put onto the moving platform to acquire image at a speed of 1.16 mm/s and the exposure time of the CCD (Charge coupled device) camera (IPX-2M30, Imperx Inc., USA) was set to 8.8 ms. The spectral range was 370–1037 nm with the nominal spectral resolution of 2.73 nm. However, the starting and ending wavelengths in the spectral region had a disturbing noise level; therefore, only the spectral range of 400–1000 nm (with 1107 bands) was chosen and utilized for analysis.\textsuperscript{[19]} To remove the dark current noise in the camera, the acquired hyperspectral images were corrected as the follows:

\[
R_c = \frac{R - B}{W - B}
\]

where \(R_c\) is the corrected image, \(R\) is the raw hyperspectral image, \(B\) is the dark image (approximately 0% reflectance) acquired by covering the camera completely with the lens cover, and \(W\) is the white image (approximately 99% reflectance) acquired by a standard white tile (the reference white material).
Selection of important wavelengths

The 200 × 200 pixels region from the centre of the corrected images was selected as the region of interest (ROI). The spectrum of each pixel contained in ROI was extracted and then averaged using ENVI (Version 4.7, ITT Visual Information Solutions, Boulder, USA), which represented the spectral data of the tea sample. The current hyperspectral images have high dimensional data, resulting in increasing computational burden. Besides, the adjacent band images are highly correlated. Therefore, it is necessary to select the dominant wavelengths related to tea categories from the hyperspectral data. PCA is one of the commonly used methods used for the dimension reduction of hyperspectral data.\(^2\) After processing, the hyperspectral image contained low dimension and high signal-to-noise ratio of data. PCA transformed the hyperspectral image into several principal component (PC) images, where each PC image was a linear combination of all of the original images. The first few PC images had a larger variance contribution rate and contained most information of the origin image. The loadings of PCA were the weighting coefficients of each wavelength image at each PC, and they can be used to select the dominant wavelength image. Then, the spectral information at the dominant wavelengths was saved as spectral features. The procedure of PCA was also carried out in ENVI 4.7.

Extraction of textural features

Grey-level co-occurrence matrix (GLCM)\(^2\) was used for extracting the textural feature from the ROI of the images at dominant wavelengths. ROIs were of 900 × 900 pixels from the centre of the images. GLCM is the matrix function obtained by calculating the specified distances and directions from the sample’s adjacent pixels. The distance equals to 1 and the four directions at angles (0°, 45°, 90°, and 135°) were applied to extract four textural features, including contrast, correlation, energy, and homogeneity. The features were calculated using a procedure in Matlab 2014a (The Mathworks Inc., Massachusetts, USA). The textural features were extracted by following Eqs. (2)–(5):

\[
\text{Contrast} = \sum_i \sum_j (i - j)^2 \cdot p(i,j)
\]

\[
\text{Correlation} = \frac{\sum_i \sum_j (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}
\]

\[
\text{Energy} = \sum_i \sum_j p(i,j)^2
\]

\[
\text{Homogeneity} = \sum_i \sum_j \frac{p(i,j)}{1 + (i - j)^2}
\]

where \(\mu_x = \sum_i \sum_j i \cdot p(i,j), \mu_y = \sum_i \sum_j j \cdot p(i,j), \sigma_x = \sqrt{\sum_i \sum_j (i - \mu_x)^2p(i,j)}, \sigma_y = \sqrt{\sum_i \sum_j (j - \mu_y)^2p(i,j)}.\)

\(p(i,j)\) is \((i, j)\) th entry of the co-occurrence probability matrix.

Combination of spectral and textural features

Generally, data fusion can be divided into three levels: pixel-level fusion, feature-level fusion, and decision-level fusion.\(^2\) Among these three levels, feature-level fusion is used to extract feature variables from spectra and images and then integrate them together, which can retain most of the original information.\(^2\) To seize as much original information as possible from the hyperspectral images and to enhance the robustness and classification ability of the model, spectral and textural
features were combined based on the feature-level fusion. While this fusion level also has the limitation that a small-value parameter would be hidden by the predictability of a larger-value parameter, even though the former may be as important as, or even more important than, the latter.\textsuperscript{[24]} The normalization procedure was used to overcome this drawback of feature-level data fusion.

**Classification models**

The classification models were developed by LDA, Lib-SVM, and ELM. LDA is a common linear classification algorithm used for pattern recognition to separate the classes of samples or objects. However, if the sample variables are highly correlated, it could affect the discrimination performance, and so PCA is used to reduce the number of variables and extract the effective features as input data to develop the LDA model.\textsuperscript{[25]} Similarly, Lib-SVM is an optimized pattern recognition algorithmic version based on SVMs, which was introduced by Chih-Chung Chang and Chih-Jen Lin.\textsuperscript{[26]} We chose the radial basis function (RBF) as the kernel function, and the parameters (c: cost, g: gamma) of RBF were optimized by the cross-validation process. ELM is proposed for training single-layer feed forward neural network (SLFN). In ELM, only the parameters between the hidden layer and the output layer need to be learned.\textsuperscript{[27]} The models were developed in Matlab 2014a software (The Mathworks Inc., Massachusetts, USA).

**Results and discussion**

**Spectral data analysis**

The mean spectral curves of tea samples within the wavelength range of 400–1000 nm are shown in Fig. 1. The general trends of the spectral curves were similar, except for that of green tea, which was a bit different from the other four teas. An absorption peak around 785 nm appeared in the spectral curves of green tea, whereas for all the spectral curves, two main absorption bands appeared at around 670 and 970 nm. According to previous research, an absorption band at 670 nm was associated with chlorophyll.\textsuperscript{[28]} The absorption around 785 nm was due to the fourth overtone of N-H vibration, and near 970 nm it was most related to water.\textsuperscript{[29]}

To reduce the spectral dimension, PCA was applied to all hyperspectral images to select the dominant wavelengths and images. The first three PC images provided 99.26% cumulative variance.

![Figure 1. The mean spectral curves of tea samples in wavelength range of 400–1000 nm.](image-url)
contribution rate. Thus, the dominant wavelength images were selected according to the loadings of the first three PC images in this study. As shown in Fig. 2, the peaks and valleys of the corresponding wavelengths at the first three PC images were selected as the dominant wavelengths. As a result, four dominant wavelengths (589, 635, 670, and 783 nm) were saved as spectral features.

**Extraction of textural features by GLCM**

Textural features were extracted by GLCM from ROI of the images at four dominant wavelengths (589, 635, 670, and 783 nm). The four dominant images at 589, 635, 670, and 783 nm are shown in Fig. 3. The textural features contrast, correlation, energy, and homogeneity were calculated at four directions (0°, 45°, 90°, and 135°). As a result, each tea sample had 64 textural features.

**Classification of tea categories**

Before developing the classification models, standard normal variate (SNV) was applied to reduce spectral noise. After spectral processing, LDA, Lib-SVM, and ELM classification models based on full spectra were developed, respectively. The classification results are presented in Table 1. Lib-SVM and ELM models achieved a high correct classification rate (CCR). However, the full spectra contained abundant information and the processed data was time consuming and unsuitable for online system. Thus, selection of the spectral features was necessary. Four dominant wavelengths (589, 635, 670, and 783 nm) were selected as spectral features by PCA. As shown in Table 1,
compared with the LDA, Lib-SVM, and ELM models based on full spectra, the CCR of the spectral features models declined from 83.87% to 70.97%, 98.39% to 75.81%, and 96.77% to 80.65%, respectively. However, the variables of the models were reduced significantly, which improved the data-processing speed.

Based on the textural features, LDA, Lib-SVM, and ELM classification models were established, respectively. The LDA and Lib-SVM models obtained the same CCR of 82.26% in the prediction set. The ELM classifier with CCR of 91.94% acquired the best result. Compared to the LDA, Lib-SVM, and ELM models built with spectral features, the models built with textural features had a relatively high performance in either the calibration set or the prediction set. It might be that the textural features were also important for classifying categories.

Based on feature-level data fusion, we combined the spectral and textural features, followed by the normalization procedure to overcome the problem that a large-value parameter would hide the predicted ability of a smaller-value parameter. With the input of data fusion, the LDA, Lib-SVM, and ELM classification models were built. All of the models’ classification results are shown in Table 1. Compared to the results of the models developed with spectral or textural features, the three models based on data fusion had higher CCR in both the calibration set and the prediction set. It is might be that the data fusion model combined both the external and internal attributes of tea, which could explain more quality of tea. The Lib-SVM model based on data fusion or full spectra had the highest CCR of 98.39%. However, the variables of the data fusion Lib-SVM model were lesser than the full spectra model. The Lib-SVM model based on data fusion had the best performance and robustness.

### Table 1. CCR (%) for the calibration set and the prediction set based on different features.

<table>
<thead>
<tr>
<th>Models</th>
<th>Types of data</th>
<th>Number of variables</th>
<th>Calibration set</th>
<th>Prediction set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td>Full spectra</td>
<td>1107</td>
<td>73.39</td>
<td>83.87</td>
</tr>
<tr>
<td></td>
<td>Spectral features</td>
<td>4</td>
<td>79.03</td>
<td>70.97</td>
</tr>
<tr>
<td></td>
<td>Textural features</td>
<td>64</td>
<td>83.87</td>
<td>82.26</td>
</tr>
<tr>
<td></td>
<td>Data fusion</td>
<td>68</td>
<td>87.10</td>
<td>83.87</td>
</tr>
<tr>
<td></td>
<td>Full spectra</td>
<td>1107</td>
<td>100</td>
<td>98.39</td>
</tr>
<tr>
<td>Lib-SVM</td>
<td>Spectral features</td>
<td>4</td>
<td>79.84</td>
<td>75.81</td>
</tr>
<tr>
<td></td>
<td>Textural features</td>
<td>64</td>
<td>94.35</td>
<td>82.26</td>
</tr>
<tr>
<td></td>
<td>Data fusion</td>
<td>68</td>
<td>100</td>
<td>98.39</td>
</tr>
<tr>
<td></td>
<td>Full spectra</td>
<td>1107</td>
<td>97.58</td>
<td>96.77</td>
</tr>
<tr>
<td>ELM</td>
<td>Spectral features</td>
<td>4</td>
<td>79.84</td>
<td>80.65</td>
</tr>
<tr>
<td></td>
<td>Textural features</td>
<td>64</td>
<td>97.58</td>
<td>91.94</td>
</tr>
<tr>
<td></td>
<td>Data fusion</td>
<td>68</td>
<td>100</td>
<td>96.77</td>
</tr>
</tbody>
</table>

Table 1.

### Model verification

To test the practicability of the model, each tea category had four samples that were not included in the calibration set and the prediction set, and were utilized for classification based on the data fusion Lib-SVM model. Classification results are illustrated in Table 2. The total CCR achieved 95%; only one – Oolong tea – was misclassified as yellow tea. The results revealed that the performance and robustness of the model were at a very high level.

### Table 2. Classification results of the data fusion Lib-SVM model in new samples.

<table>
<thead>
<tr>
<th>Tea categories</th>
<th>Samples number</th>
<th>Classification result</th>
<th>Total correct classification rate ( % )</th>
</tr>
</thead>
<tbody>
<tr>
<td>White tea</td>
<td>4</td>
<td>White tea 4, Black tea 0, Yellow tea 0, Green tea 0, Oolong tea 0</td>
<td>95</td>
</tr>
<tr>
<td>Black tea</td>
<td>4</td>
<td>White tea 0, Black tea 4, Yellow tea 0, Green tea 0, Oolong tea 0</td>
<td></td>
</tr>
<tr>
<td>Yellow tea</td>
<td>4</td>
<td>White tea 0, Black tea 0, Yellow tea 4, Green tea 0, Oolong tea 0</td>
<td></td>
</tr>
<tr>
<td>Green tea</td>
<td>4</td>
<td>White tea 0, Black tea 0, Yellow tea 0, Green tea 4, Oolong tea 0</td>
<td></td>
</tr>
<tr>
<td>Oolong tea</td>
<td>4</td>
<td>White tea 0, Black tea 0, Yellow tea 1, Green tea 0, Oolong tea 3</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

The results suggested that visible and NIR hyperspectral imaging combined with data fusion could be used to classify the tea categories. The highest CCR obtained from the Lib-SVM model based on data fusion achieved 98.39%, which were superior to the models based on single features. Future research should focus on data fusion for the classification or prediction of the study object, and optimization of data processing. Furthermore, much more efforts are needed to include more samples to improve the performance of the classifying models, and explore more chemometrics algorithm, such as partial least square discriminant analysis (PLS-DA) and Random forest (RF).

Acknowledgements

This work was supported by the National key Research & Development Plan (2016YFD0200900) and Modern Agriculture (tea) Special System of China (CARS-23). The authors thank Professor Chen for assisting in modification.

References